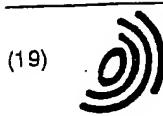


English equivalent of (1)



(19) Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) EP 0 717 357 B1

(12)

EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention
of the grant of the patent:
26.02.2003 Bulletin 2003/09

(51) Int Cl.7: G06F 11/10, G11B 20/18

(21) Application number: 95309117.0

(22) Date of filing: 14.12.1995

(54) Disk array apparatus

Speicherplattenanordnungsgerät
Appareil de réseau de disques

(84) Designated Contracting States:
DE FR GB

(30) Priority: 16.12.1994 US 357847

(43) Date of publication of application:
19.06.1996 Bulletin 1996/25

(73) Proprietor: HYUNDAI ELECTRONICS AMERICA
Milpitas, California 95035 (US)

(72) Inventors:

- Stewart, John W.
Wichita, KS 67232 (US)
- Gates, Dennis E.
Wichita, KS 67220 (US)

- Dekoning, Rodney A.
Wichita, KS 67226 (US)
- Rink, Curtis W.
Wichita, KS 67205 (US)

(74) Representative: Gill, David Alan
W.P. Thompson & Co.,
Celon House,
289-293 High Holborn
London WC1V 7HU (GB)

(56) References cited:
EP-A- 0 369 707 US-A- 5 257 391
US-A- 5 373 512

EP 0 717 357 B1

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

Description

[0001] The present invention relates to disk array apparatus and in particular, but not exclusively, to a hardware disk array controller providing improved performance, and the execution of concurrent data movements and parity calculations.

[0002] A Redundant Array of Inexpensive Disks, referred to as a RAID storage system, is a collection of disk drives which appears as a single large disk drive to a host computer system. Additionally, part of the disk storage capacity is utilized to store redundant information about user data stored on the remainder of the storage capacity. This redundant information allows the disk array to continue to function without the loss of data should an array disk drive member fail, and permits the regeneration of data to a replacement array disk drive member.

[0003] Several RAID disk array design alternatives are presented in an article titled "A Case for Redundant Arrays of Inexpensive Disks (RAID)" by David A. Patterson, Garth Gibson and Randy H. Katz; University of California Report No. UCB/CSD 87/391, December 1987. The article discusses disk arrays and the improvements in performance, reliability, power consumption and scalability that disk arrays provide in comparison to single large magnetic disks.

[0004] Five disk array arrangements are described in the article. The first level RAID comprises N disks for storing data and N additional "mirror" disks for storing copies of the information written to the data disks. RAID level 1 write functions require that data be written to two disks, the second "mirror" disk receiving redundant information, i.e., the same information provided to the first disk. When data is read, it can be read from either disk.

[0005] RAID level 3 systems comprise one or more groups of N+1 disks. Within each group, N disks are used to store data, and the additional disk is utilized to store redundant information, i.e., parity information. During RAID level 3 write functions, each block of data is divided into N portions for storage among the N data disks. The corresponding parity information is calculated by determining the exclusive-OR product of the data written to the N data disks and written to a dedicated parity disk. When data is read, all N data disks must be accessed. The parity disk is used to reconstruct information in the event of a disk failure.

[0006] A RAID level 2 system is similar to the RAID level 3 systems described above, but includes additional redundant disks for identifying disk drive failures.

[0007] RAID level 4 systems are also comprised of one or more groups of N+1 disks wherein N disks are used to store data, and the additional disk is utilized to store parity information. RAID level 4 systems differ from RAID level 3 systems in that data to be saved is divided into larger portions, consisting of one or many blocks of data, for storage among the disks. Write operations still require access to two disks, i.e., one of the N data disks

and the parity disk. In a similar fashion, read operations typically need only access a single one of the N data disks, unless the data to be read exceeds the block length stored on each disk. As with RAID level 3 systems, the parity disk is used to reconstruct information in the event of a disk failure.

[0008] RAID level 5 is similar to RAID level 4 except that parity information, in addition to the data, is distributed across the N+1 disks in each group. Although each group contains N+1 disks, each disk includes some blocks for storing data and some blocks for storing parity information. The location at which parity information is stored is controlled by an algorithm implemented by the user. As in RAID level 4 systems, RAID level 5 write operations require access to at least two disks. However, no longer does every write operation to a group require access to the same dedicated parity disk, as in RAID level 4 systems. This feature provides the opportunity to perform concurrent write operations.

[0009] As with RAID level 3, parity data in either a RAID level 4 or 5 system can be calculated by performing a bit-wise exclusive-OR of corresponding portions of the data stored across the N data drives. However, because each parity bit is simply the exclusive-OR product of all the corresponding data bits from the data drives, new parity can be more easily determined from the old data and the old parity as well as the new data in accordance with the following equation:

$$30 \text{ new parity} = (\text{old data XOR new data}) \text{ XOR old parity.}$$

[0010] Although the parity calculation for RAID levels 4 or 5 shown in the above equation is much simpler than performing a bit-wise exclusive-OR of corresponding portions of the data stored across all of the data drives, a typical RAID level 4 or 5 write operation will require a minimum of two disk read and two disk write operations. More than two disk read and write operations are required for data write operations involving more than one data block. Each individual disk read operation involves a seek and rotation to the appropriate disk track and sector to be read. The seek time for all disks is therefore the maximum of the seek times of each disk. A RAID level 4 or 5 system thus carries a significant write penalty when compared with a single disk storage device or with RAID level 1, 2 or 3 systems.

[0011] In order to coordinate the operation of the disk drives within an array to perform read and write functions, to map received data onto the array disk drive members, generate and check redundant information, and provide data restoration and reconstruction, complex storage management techniques are required. In many early disk array systems, the array management software necessary to perform these complex storage management techniques is executed within the host computer system. The host system thereby functions as the disk array controller and performs the generation

and checking of redundant information as well as coordinating the many other storage management operations required of the disk array. Having the host perform these functions is expensive in host processing overhead.

[0012] The present invention seeks to provide for an improved disk array apparatus and also a related method of operating such arrays.

[0013] In accordance with the present invention there is provided disk array controller apparatus for operation with a host computer and a plurality of disk drive means, comprising a first high speed local bus, an interface connecting said first high speed local bus with said host computer, at least one interface connecting said plurality of disk drive means to said first high speed local bus and processor means connected to said first high speed local bus for controlling the operation of the components connected to said first high speed local bus, characterized by parity assist logic means connected directly to said first high speed local bus, local data storage means connected to said first high speed local bus, and a second high speed local bus connecting said parity assist logic means to said local data storage means wherein said parity assist logic means is arranged to provide data via said second high speed local bus to said local data storage means for the determination of parity data during disk array write operation.

[0014] The present invention advantageously provides a new and useful disk array controller hardware architecture that improves controller performance.

[0015] Also, the present invention provides a disk array controller architecture which allows concurrent data movements and parity calculations, and the concurrent execution of time independent tasks.

[0016] It is yet another advantage of the present invention that a disk array controller can be provided which includes unique parity assist logic and a dedicated local memory for performing parity generation and data reconstruction operations.

[0017] Further the present invention advantageously can provide a new and useful disk array controller architecture which permits the simultaneous operation of data block moves between storage I/O devices and a local memory; data block moves between a host system and the local memory; parity calculations; and normal array controller processor memory fetches.

[0018] It is yet another advantage of the present invention that a disk array controller architecture which also permits queued operation of block moves and queued operation for parity tasks can be provided.

[0019] It should also be appreciated that the invention relates to the disk array control method embodied within the above apparatus.

[0020] Preferably there can be provided, in accordance with the present invention, a disk array system for storing data received from, and providing stored data to, a host computer system. The disk array system includes a first high speed local bus; an interface circuit connect-

ing the first high speed local bus with the host computer system; a plurality of disk drive members; at least one interface circuit connecting the plurality of disk drive members with the first high speed local bus; a parity assist logic circuit connected to the first high speed local bus; a local memory storage; a second high speed local bus connecting the parity assist logic with said local memory storage; and a processor connected to the first high speed local bus for controlling the operation of the components connected to said first high speed local bus. The parity assist logic provides data via the second high speed local bus to the local memory required for the calculation of parity data during disk array write operations, and manipulates the data saved to the local memory to determine parity during disk array write operations.

[0021] The invention can also provide a disk array control apparatus for a disk array system including a plurality of disk drive members for storing data received from, and providing stored data to, a host computer system, said disk array controller comprising a first high speed local bus, an interface circuit connecting said first high speed local bus with said host computer system, at least one interface circuit connecting said plurality of disk drive members with said first high speed local bus, a parity assist logic circuit connected to said first high speed local bus, a local memory storage, a second high speed local bus connecting said parity assist logic with said local memory storage, wherein said parity assist logic provides data via said second high speed local bus to said local memory required for the calculation of parity data during disk array write operations, and manipulates the data saved to said local memory to determine parity during said disk array write operations, and a processor connected to said first high speed local bus for controlling the operation of the components connected to said first high speed local bus.

[0022] In the described embodiment, the parity assist logic includes a parity assist engine, including exclusive-OR logic, for calculating parity; a dual ported memory interface connecting said parity assist engine with said local memory via said second high speed local bus; and an interface circuit connecting said dual ported memory interface with said first high speed local bus.

[0023] The first local bus is a high speed PCI buses. The second bus is a dual port memory interface.

[0024] The array controller architecture is scalable and supports RAID modes 0, 3, 4 and 5. The architecture features a high bandwidth parity calculation engine and a buffered PCI interface operating at the full speed of the high speed PCI local bus. It features multi processing so that multiple tasks may be queued for execution. Additionally, it provides a way to attach blocks of up to 128 MBytes of additional local RAM memory to the PCI bus. The additional local memory is dual ported so that PCI and parity operations may operate concurrently.

[0025] Thus, it should be appreciated that the inven-

tion can advantageously provide for a high performance scalable hardware architecture for a disk array storage subsystem which supports RAID modes 0, 3, 4 and 5. The architecture features a high bandwidth parity calculation engine, a buffered PCI interface operating at the full speed of a PCI bus, and a dedicated local memory. The dedicated local memory is dual ported so that PCI and parity operations may operate concurrently. The architecture of the disk array controller allows parity calculations and memory block moves to occur without interfering with the controller processor or its associated memory, freeing the controller processor to manage array task control. The array controller configuration allows simultaneous operation of data block moves between storage I/O devices and local memory; data block moves between host SCSI connections and local memory; parity calculations; and normal CPU memory fetches, queued operations for block moves and queued operations for parity tasks.

[0025] The invention is described further hereinafter by way of example only, with reference to the accompanying drawings in which:

Fig. 1 is a simple block diagram of a prior art disk array system utilizing a high speed PCI bus; and Fig. 2 is a simple block diagram of a disk array system utilizing a high speed PCI bus and a high performance parity function architecture in accordance with an embodiment of the present invention.

[0026] Most disk array systems in use today are self-contained, including a dedicated controller for executing the array management software, thus relieving the host system of these operations. A simple block diagram of a known disk array system is shown in Fig. 1. The system includes an intelligent array controller 100 for managing the transfer of data between a host computer system 12 and N disk drive units, five of which, identified as DRIVE A through DRIVE E, are shown in Fig. 1. Central to the array controller is a high speed local bus 102, such as a Peripheral Component Interconnect (PCI). A host SCSI interface 104 and SCSI bus 14 provide connection between the host computer system 12 and PCI bus 102. Similarly, each of disk drives DRIVE A through DRIVE E are connected to PCI bus 102 through a SCSI drive interface, identified by reference numerals 112A through 112E, respectively, and corresponding SCSI buses 114A through 114E. Parity functions are performed by a parity logic circuit 108 and local memory 110, both of which are also connected to the PCI bus 102. Communication between, and operation of, the controller components is controlled by a processor 106, in accordance with instructions residing in a processor memory 118. The construction and operation of the array controller shown in Fig. 1, as well as the components included in the controller, will be readily understood by those skilled in the art.

[0027] The RAID storage process requires many par-

ity calculations and data movement operations to create the necessary data redundancy, or reconstruct data following a disk failure. In the array controller architecture shown in Fig. 1 and described above, much use of the PCI bus 102 is required to transfer new data, old data, reconstructed data, old parity information and new parity information between the host computer system 12, The array drives DRIVE A through DRIVE E, the local memory 110 and the parity logic 108 to generate new parity information during an array write operation or to reconstruct data following an array failure.

[0028] The disk array system shown in Fig. 2 includes many of the components shown in Fig. 1 and discussed earlier. For those components common to the controller architectures and array systems, the reference numerals of Fig. 1 have been retained in Fig. 2. Common components of the systems shown in Figs. 1 and 2 include: high speed PCI bus 102, host SCSI interface 104, SCSI bus 14, SCSI drive interfaces 112A through 112E, SCSI buses 114A through 114E, disk drive members DRIVE A through DRIVE E, processor 106 and processor memory 118.

[0029] However, the array controller architecture shown in Fig. 2 improves upon previously known array controller architectures by removing the parity logic and the local memory utilized for parity generation and the reconstruction of data from the PCI bus 102. The controller architecture shown in Fig. 2 provides an independent parity assist engine 132 and local memory 136 interconnected through a dual ported memory interface 134. A high speed PCI interface 130 provides for the transmission of data between PCI bus 102 and local memory 136.

[0030] The PCI bus interface 130, parity calculation engine 132, and dual ported memory interface 134 are shown integrated together into a single RAID parity assist chip 128. However, these three components operate independently, in a coordinated manner, to provide high speed data storage and parity calculations while using a minimum of host CPU cycles. A high speed local bus 138 connects the RAID parity assist chip with the local memory 136.

[0031] During a disk write operation, data blocks are received from the host SCSI channel 14 and stored in local memory 136. The parity assist engine 132 reads the data, calculates the parity information and writes it back to local memory 136. The final data blocks are then written to the appropriate disk drives from the local memory 136.

[0032] Disk read operations start with data moved from the disk drives to local memory 136. This data is then transferred to the host SCSI channel 14 unless the parity information is needed to reconstruct the data.

[0033] Since this architecture is very memory intensive, local memory 136 is organized for high bandwidth operation with a 72 bit wide interface. It also supports interleaved cycles for maximum performance. The memory system is capable of operating at the full clock

speed of the RAID parity assist chip. Refresh control is also included so that economical dynamic RAM can be used for storage. A powered-down mode of operation is provided to protect the memory data if power fails.

[0034] The PCI interface module 130 is capable of full speed PCI bus transfers. It is designed to respond to all bus modes from single byte to full bandwidth bursts. The PCI module contains a 128 byte (First-In First-Out) FIFO register to store data accepted from the PCI bus. This allows the PCI transfer to continue even if local memory is temporarily unavailable.

[0035] The parity assist engine 132 operates on data stored in the local memory 136. It calculates parity a burst at a time utilizing a 128 byte FIFO to store the intermediate results. The intermediate results are not written back to local memory. For maximum performance the control logic for the parity assist engine maintains pointers to each needed data block. The parity engine operates at the full speed of the local memory bus to provide the fastest possible performance for the memory bandwidth.

[0036] The parity assist engine contains 4 separate sections that allow additional tasks to be queued up while one task executes. Each task maintains its own control and status register so that task scheduling does not interfere with the currently executing task. Several configuration options are provided to tailor the parity engine to the array organization. The engine can be configured as a single engine which works on wide arrays up to 22+1 drives wide, or as a four engine machine which operates with arrays as wide as 4+1 drives. An intermediate mode provides a two engine machine with up to 10+1 drives.

[0037] The parity engine includes exclusive-OR logic providing RAID 5 and RAID 3 parity generation/checking as well as move and zero check modes.

[0038] One of the most important parts of the architecture is the time independent operation it provides. Blocks of data are not required to be accessed simultaneously for parity calculations. Disk operations which span several drives may be scheduled and executed as soon as possible by each device. Unrelated disk operations may continue even though all drive operations for a single task are not yet complete. This independence improves the performance of the slowest part of the system, the disk drive. It also simplifies the software task of managing concurrent hardware resource requirements.

[0039] The invention features time independent data block moves; time independent parity calculations; full queued RAID operations (Up to 3 queued tasks); full speed PCI transfer operations (4 bytes/clock cycle); and full speed parity calculations (4 bytes/clock cycle). Furthermore, the architecture is scalable via the addition of additional RAID Parity Assist chips and local memory storage.

[0040] It can thus be seen that there has been provided by the present invention a disk array controller archi-

ecture having an independent memory structure for buffering data, and an independent parity calculation engine permitting parallel operation of array tasks. It allows parity calculations and memory block moves to occur without interfering with the CPU or its memory, freeing the CPU to manage array task control. The disclosed array controller configuration allows simultaneous operation of data block moves between storage I/O devices and local memory; data block moves between host SCSI connections and local memory; parity calculations; and normal CPU memory fetches, queued operations for block moves and queued operations for parity tasks.

[0041] It will therefore be appreciated that the invention can provide for a disk array system including a first high speed local bus; an interface circuit connecting said first high speed local bus with said host computer system; a plurality of disk drive members; at least one interface circuit connecting said plurality of disk drive members with said first high speed local bus; and a processor connected to said first high speed local bus for controlling the operation of the components connected to said first high speed local bus; having the improvement comprising a parity assist logic circuit connected to said first high speed local bus; a local memory storage; and a second high speed local bus connecting said parity assist logic with said local memory storage; wherein said parity assist logic provides data via said second high speed local bus to said local memory required for the calculation of parity data during disk array write operations, and manipulates the data saved to said local memory to determine parity during said array write operations.

[0042] The invention is not restricted to the details of the foregoing embodiment and it should be understood that various changes may be made within the scope of the appended claims.

Claims

1. Disk array controller apparatus for operation with a host computer (12) and a plurality of disk drive means (114A-114E), comprising a first high speed local bus (102), an interface (104) connecting said first high speed local bus (102) with said host computer (12), at least one interface (112A-112E) connecting said plurality of disk drive means (114A-114E) to said first high speed local bus (102) and processor means (106) connected to said first high speed local bus (102) for controlling the operation of the components connected to said first high speed local bus (102), characterized by parity assist logic means (132) connected directly to said first high speed local bus (102), local data storage means (136) connected to said first high speed local bus (102), and a second high speed local bus (138) connecting said parity assist logic means (132) to said local data storage means (136), wherein said

parity assist logic means (132) is arranged to provide data via said second high speed local bus (138) to said local data storage means (136) for the determination of parity data during disk array write operation.

2. Apparatus as claimed in Claim 1, wherein said parity assist logic means (132) comprises a parity assist engine (132), a dual ported memory interface (134) connecting said parity assist engine (132) with said local data storage means (136) via said second high speed local bus (138), and an interface circuit (130) for connecting said dual ported memory interface (134) with said first high speed local bus (102).
3. Apparatus as claimed in Claim 2, wherein said parity assist engine (132) comprises an exclusive-OR circuit.

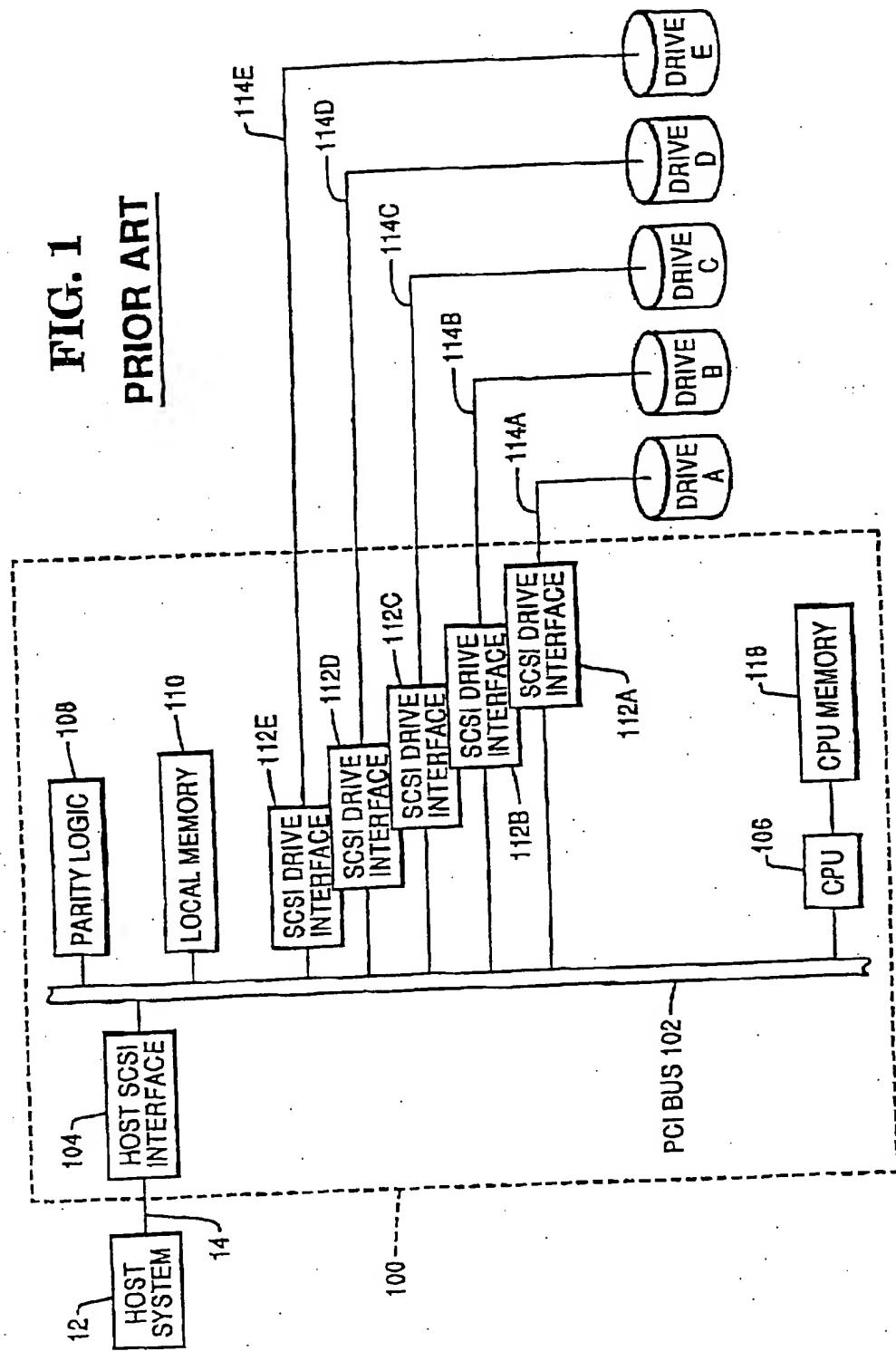
Revendications

1. Appareil contrôleur de grappe de disques destiné à fonctionner avec un ordinateur hôte (12) et une pluralité de moyens d'entraînement de disques (114A-114E), comprenant un premier bus local à haute vitesse (102), une interface (104), connectant ledit premier bus local à haute vitesse (102) audit ordinateur hôte (12), au moins une interface (112A-112E) connectant ladite pluralité de moyens d'entraînement de disques (114A-114E) audit premier bus local à haute vitesse (102) et un moyen processeur (106) connecté audit premier bus à haute vitesse, pour contrôler le fonctionnement des composants connectés audit premier bus local à haute vitesse (102), caractérisé par un moyen logique d'aide de parité (132) connecté directement audit premier bus local à haute vitesse (102), un moyen de stockage de données local (136) connecté audit premier bus local à haute vitesse (102), et un deuxième bus local à haute vitesse (138) connectant ledit moyen logique d'aide de parité (132) audit moyen de stockage de données local (136), où ledit moyen logique d'aide de parité (132) est安排 pour fournir des données par l'intermédiaire dudit deuxième bus local à haute vitesse (138), audit moyen de stockage de données local (136), pour déterminer des données de parité pendant une opération d'écriture de grappe de disques.
2. Appareil tel que revendiqué dans la revendication 1, dans lequel ledit moyen logique d'aide de parité (132) comprend un moteur d'aide de parité (132), une interface de mémoire à double accès (134) connectant ledit moteur d'aide de parité (132) audit moyen de stockage de données local (136), par ledit deuxième bus local à haute vitesse (138), et un circuit d'interface (130) pour connecter ladite interface de mémoire à double accès (134) audit premier bus local à haute vitesse (102).
3. Appareil tel que revendiqué dans la revendication 2, dans lequel ledit moteur d'aide de parité (132) comprend un circuit OU exclusif.

10 Patentansprüche

1. Plattenarray-Controller für den Betrieb mit einem Hostcomputer (12) und einer Mehrzahl von Plattenlaufwerken (114A-114E), umfassend einen ersten schnellen lokalen Bus (102), eine Schnittstelle (104), die den genannten ersten schnellen lokalen Bus (102) mit dem genannten Hostcomputer (12) verbindet, wenigstens eine Schnittstelle (112A-112E), die die genannte Mehrzahl von Plattenlaufwerken (114A-114E) mit dem genannten ersten schnellen lokalen Bus (102) verbindet, und einen Prozessor (106), der mit dem genannten ersten schnellen lokalen Bus (102) verbunden ist, um den Betrieb der mit dem genannten ersten schnellen lokalen Bus (102) verbundenen Komponenten zu steuern, gekennzeichnet durch Paritätsassist-Logik (132), die direkt mit dem genannten ersten schnellen lokalen Bus (102) verbunden ist, ein lokales Datenspeichermittel (136), das mit dem genannten ersten schnellen lokalen Bus (102) verbunden ist, und einen zweiten schnellen lokalen Bus (138), der die genannte Paritätsassist-Logik (132) mit dem genannten lokalen Datenspeichermittel (136) verbindet, wobei die genannte Paritätsassist-Logik (132) die Aufgabe hat, Daten über den genannten zweiten schnellen lokalen Bus (138) zu dem genannten lokalen Datenspeichermittel (136) zu übertragen, um Paritätsdaten während des Plattenarray-Schreibvorgangs zu bestimmen.
2. Vorrichtung nach Anspruch 1, bei der die genannte Paritätsassist-Logik (132) eine Paritätsassist-Maschine (132), eine Dualport-Speicherschnittstelle (134), die die genannte Paritätsassist-Maschine (132) mit dem genannten lokalen Datenspeichermittel (136) über den genannten zweiten schnellen lokalen Bus (138) verbindet, und eine Schnittstellschaltung (130) zum Verbinden der genannten Dualport-Speicherschnittstelle (134) mit dem genannten ersten schnellen lokalen Bus (102) umfasst.
3. Vorrichtung nach Anspruch 1, bei der die genannte Paritätsassist-Logik (132) eine Exklusiv-ODER-Schaltung umfasst.

FIG. 1
PRIOR ART



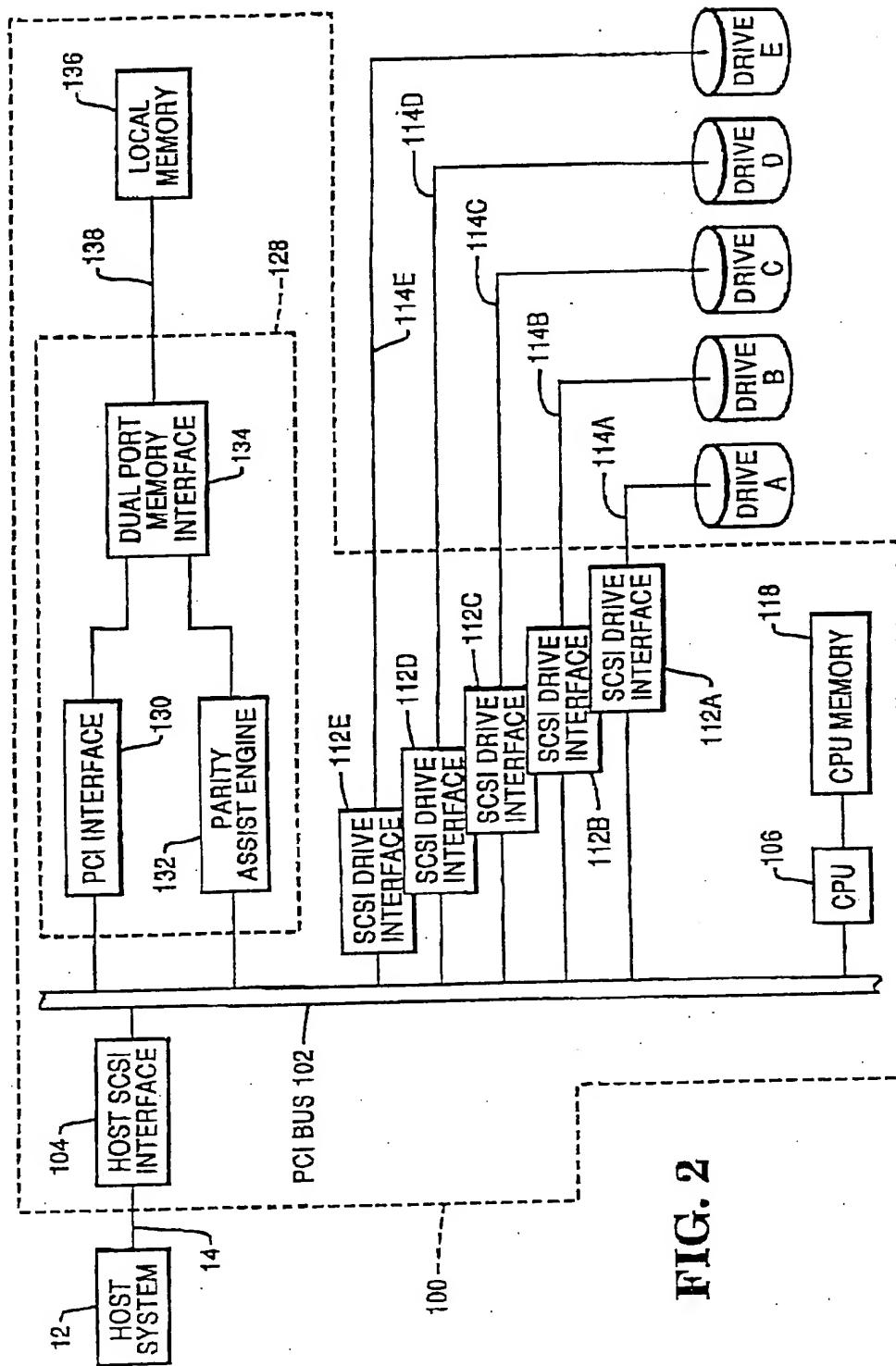


FIG. 2